Accepted on 1 November 2019 by the *European Journal of Politics and Gender*. Sent to ORCA on 19 December 2019.

Title:

Policy evaluation and gender mainstreaming in the European Union: The perfect (mis)match?

Authors:

Lut Mergaert (Yellow Window)

Rachel Minto (Cardiff University)

María Bustelo (Complutense University of Madrid)

Abstract

This article assesses the ability of the European Commission's current approach to policy evaluation to evaluate gender mainstreaming and, in turn, other cross-cutting social agendas (Articles 8-10 TFEU). Taking EU research policy as a case study, through our analysis, we reveal mismatches between current evaluation standards (adopted within the Better Regulation framework) and requirements for effectively assessing progress towards cross-cutting social objectives, such as gender equality. The article concludes with a series of recommendations to overcome the identified shortcomings. Our analysis constitutes a key contribution for the development of feminist scholarship on the post-implementation phase of the policy process.

Six key words: evaluation, gender mainstreaming, social objectives, research policy, European Commission, Better Regulation

Four key messages:

- The European Commission's Better Regulation framework provides standards for evaluation.

- The Better Regulation framework does not accommodate requirements for evaluating social agendas.

- To meet gender mainstreaming commitments, the Commission's approach to evaluation ought to change.

- The article offers suggestions for strengthening evaluation from a gender equality perspective.

Introduction

This article assesses the suitability of the European Union's (EU) dominant evaluation model to support progress towards its constitutionalised cross-cutting agendas (Articles 8-10 TFEU), taking gender mainstreaming in EU research policy as a case study. Policy evaluation is now an established feature of EU governance, with impact assessments and post-adoption evaluations enjoying an institutionalised position. Its status has been reinforced in recent years, with the launch of the Better Regulation Agenda in 2015 (revised in 2017), which aimed to streamline policy-making, tighten the policy cycle and demonstrate EU added value, with evaluation as a key component. Despite the heightened proceduralisation of evaluation in European policy-making, little is known about whether the European Commission's (EC) evaluation practice supports the EU's Treaty-based commitment to promoting cross-cutting social objectives. This article seeks to fill this gap by assessing the EC's current approach to evaluation against the distinct requirements of feminist approaches to evaluation. In turn, analysis identifies the steps necessary to improve the EC's practice of evaluation so it supports the promotion of cross-cutting social objectives, including gender equality.

Following a section on theory and methods (section 1), our article is organised around two phases of analysis: the first on evaluation in theory (section 2); the second on evaluation in practice (section 3). In section 2, we theorise the ability of the EC's approach to evaluation to

evaluate gender mainstreaming, identifying the (mis-)matches between the Commission's more rationalist approach to policy evaluation and the requirements for evaluating gender equality and gender mainstreaming. In this theoretical assessment, we turn to research on rationalist approaches to evidence-based policy-making, as the foundation of policy evaluation in the EU. Then, drawing on critical scholarship on evaluation, we assess the ability of such rationalist models to accommodate the requirements of more complex social agendas, like gender equality and mainstreaming.

Section 3 also identifies (mis-)matches, but is focused on evaluation in practice. This text-based, empirical analysis consists of two parts: the first focusing on the 13 individual evaluations undertaken during our period of study; and the second looking longitudinally at how gender equality is addressed in the evaluations over time and the overarching organising frameworks. Synthesising the findings from the preceding analyses, the next section presents a full account of the mismatches identified in both theory and practice. To conclude, we situate our research findings within the wider feminist scholarship, present a future research agenda and advance recommendations to strengthen policy evaluation in the EC as a tool to evaluate gender mainstreaming. We suggest that our findings are applicable beyond gender mainstreaming, specifically for the evaluation of other cross-cutting agendas (including the horizontal social clause [Article 9 TFEU] and non-discrimination [Article 10 TFEU]).

In this article, we show that the Commission's dominant approach to policy evaluation favours both results (or 'outcome') focused and quantitative analyses, as part of a closed policy cycle that makes no provision for longitudinal analysis. We challenge the Commission's approach as inadequate to evaluate gender mainstreaming and other cross-cutting agendas, as these demand attention to process (e.g. to the functioning of mechanisms and tools), the adoption of a longitudinal approach, and mixed methods analyses that include both quantitative and qualitative data. Notwithstanding its potential to provide a coherent framework for evaluation, the Better Regulation Agenda falls short in ensuring that gender equality concerns are addressed consistently and effectively across all evaluation exercises.

1) Theory and methods

The plethora of studies on gender mainstreaming in the EU review and critique this potentially transformative process as a means to promote gender equality across all European activity and throughout all stages of the policy cycle (Allwood, 2013; Braithwaite, 2000; de Gregorio Hurtado, 2017; Debusscher, 2011; Debusscher & Van der Vleuten, 2012; Hubert, 2012; Kronsell, 2012, 2015; Prügl, 2012; True, 2009). Within this research from Feminist Policy Studies and Gender and Politics scholars, attention has recently turned to post-adoption stages. This attention is long overdue. Within the developing field of post-adoption scholarship, early studies have concentrated specifically on implementation (Callerstig, 2014; R. Cavaghan, 2017; Engeli & Mazur, 2018; Krizsan & Roggeband, 2018; Mazur, 2017; Mergaert, 2012; Van der Vleuten, 2016), including interrogating opposition to gender equality within this phase (Verloo, 2018). Such is the interest in post-adoption research as a rich site for feminist inquiry, a research group and network have already been generated (the Gender Equality Policy in Practice Project, GEPP).. However, within this rapidly developing field, policy evaluation as a post-adoption phase of the policy process is yet to receive much considered attention. Indeed, only a couple of studies address the evaluation of gender mainstreaming in practice (Bustelo, 2003; Moser, 2005), with others reflecting on how Critical Frame Analysis can contribute to both evaluating policy formulation and reconstructing theories of change from a different epistemological point of view (Bustelo & Verloo, 2009), and the need to perform evaluation from a gender+ perspective to (re)gender the policy-making process (Bustelo, 2017). Beyond this, research is lacking. Our study seeks to contribute to this under-developed area of post-adoption scholarship, through assessing the ability of the Commission's current approach to policy evaluation to evaluate gender mainstreaming and, in turn, to promote gender equality.

The paucity of feminist research into policy evaluation is striking given the well-established position of ex-ante evaluations (i.e. impact assessments undertaken 'before the event') and ex post evaluations (i.e. post-implementation evaluations undertaken 'after the event') as part of European policy-making. The Commission's most recent activity in this area is located within its Better Regulation Agenda. It is within this framework that the EC consolidated its approach to evaluation and solidified the position of evaluation within the governance architecture of the EU. Through a set of policy instruments, including the Better Regulation Guidelines and Toolbox, the Better Regulation Agenda aimed to ensure more efficient and effective governance, including through evaluation. This included ex-ante (specifically the Integrated Impact Assessment [IIA]) and ex-post evaluations. In this article, although ex-ante evaluation

is not excluded, our focus is ex-post. Despite the EC's overall intention to streamline policy-making, the process of evaluation remains a complex one, contending with competing objectives, a multiplicity of actors and different types of knowledge. Therefore, it is a challenging site for feminist analysis but essential to gain a fuller understanding of gender mainstreaming throughout the policy process.

We adopt a case study approach, focusing on gender mainstreaming in EU research policy. This case offers a rich site for empirical analysis. Notably, it serves as a 'hypothesis-generating case study' (Pollack & Hafner-Burton, 2000), providing valuable initial findings to inform more concrete hypotheses for feminist and gender-responsive analyses in future. There is a long and (relatively) strong history of both gender mainstreaming and evaluation in this policy area (Mergaert & Minto, 2015). Even from the early days of gender mainstreaming, the Commission's gender mainstreaming activity in this area was seen as an example of better practice (Pollack & Hafner-Burton, 2000). Though rather scarce, later scholarship has been more critical, drawing attention to shortcomings in the gender mainstreaming approach (Abels, 2012; R. Cavaghan, 2013, 2017; R. M. Cavaghan, 2012; Linková & Červinková, 2011; Mergaert, 2012; Mergaert & Demuynck, 2011). Amongst this policy-focused scholarship, only more recent research has focused on the evaluation phase (Mergaert & Minto, 2015), highlighting the weak links between ex-ante and ex-post evaluations.

Given the expenditure attached to EU research policy, there is also a long-established culture of evaluation (Højlund, 2014a). Policy-making works on a cyclical basis, with evaluations built into the process of reiterating the Research and Innovation Framework Programmes (FPs). To date, there have been eight framework programmes. The first (FP1) was launched in 1984, with a five-year budget of ECU 3.75 billion and a clear industrial focus. The current framework programme (Horizon 2020) is significantly broader in scope, spanning seven years, with a budget of nearly €80 billion.[1] From FP1 to Horizon 2020, the cyclical policy-making process has been repeated seven times, although it was not until FP5 (1998-2002) that gender was addressed. From this time, gender equality objectives have been included within the legal basis

[1] These figures are not directly comparable as the overall budget and component allocations span different time periods (five years for FP1, seven for Horizon 2020 and the component Euratom continues to function on a five-year cycle).

underpinning EU research policy; and, since FP7, all framework programmes have been situated within the wider European Research Area which has gender equality as a central priority (European Commission, 2012, p. 4). It is here that the analysis begins: from FP5 (1998-2002), over FP6 (2002-2006), FP7 (2007-2013), up to Horizon 2020 (2014-2020).

2) Contrasting conventional and critical approaches to evaluation

The rise of policy evaluation in the Commission has taken place within a 'new public management' context, which has supported and characterised evaluation's widespread adoption (Minto & Mergaert, 2018). A key feature of new public management is its preference for rationalist, evidence-based policy-making, which understands the policy cycle as a closed loop, distinguishing between various stages of the policy process, commonly: agenda setting, formulation, decision-making, implementation, and evaluation (e.g. Versluis, van Keulen, & Stephenson, 2011). In this model, evaluation sets out to have an accountability and policy learning role as part of effective and efficient governance. It is used to measure policy output and determine the 'added value' of policy. This closed-loop, compartmentalised model of the policy process – where evaluations measure and record policy-associated change – characterises the Commission's approach to evaluation.

There is wide recognition that the EC has been a significant promoter of evaluation in Europe (Stern, 2009; Summa & Toulemonde, 2002), with an exponential growth in evaluation activity in the EU from the late 1990s, across policy domains and phases of the policy cycle. Ensuring financial accountability has underpinned development, with ex-post evaluation enjoying a strong legal and political basis in the EU's Financial Regulations and associated sector-specific regulations (e.g. in the Structural Funds). The 1996 Communication on Evaluation (European Commission, 1996) required all Directorates-General (DGs) to have their own evaluation functions or units. Following this, the Communications in 2010 (European Commission, 2010) and 2013 (European Commission, 2013) embedded evaluation within the Smart Regulation framework (the predecessor to Better Regulation); although nearly a decade prior to that the 2002 Communication had integrated evaluation within the decision-making process and linked it to the EU's policy cycle. From this relatively early stage, the Commission committed to analysing intervention logics. These concern cause and effect relationships linking how an

intervention attains its objectives, which tend to favour rationalistic and causal approaches to evaluation. The Commission's emphasis on accountability further compounded the rationalist approach, with 'a tendency to formally favour indicators and quantification where possible' (Stern, 2009, p. 71).

Despite the dominance of the Commission's rationalist approach, the Better Regulation Toolbox explicitly recognises the complexity of evaluating EU interventions (European Commission, 2017b, pp. 269-270):

> '(…) "cause and effect" relationships are challenging to prove, particularly when evaluating EU policies which operate in a complex environment influenced by a wide range of factors falling outside the scope of the EU intervention. When evaluating EU legislation, it is particularly difficult to identify a robust counter-factual situation (i.e. what the situation would be if EU laws had not been adopted), making absolute quantitative analysis problematic.'

Notwithstanding this acknowledgement, the Commission's dual emphasis upon 'proportionality' and 'simplification' serves to flatten this complexity and, in turn, risks a continued recourse to quantitative data to verify the performance of interventions. This risk is compounded by the inclusion in the Better Regulation Toolbox (2017) of a whole chapter (chapter 8) on quantitative methods and data, without an equivalent chapter on qualitative or mixed methods approaches. Furthermore, and despite the arguments recognising the difficulties and complexity for 'causal' evaluation, there is an explicit methodological preference and hierarchy that present this causal evaluation as the 'gold standard'. This is the picture painted in the most recent Better Regulation Guidelines, which state that 'when causal evaluation is not possible, or only at disproportionate effort, EU evaluations have to rely on qualitative, reasoned arguments (backed by the appropriate quantitative and qualitative evidence) about the likely role/contribution of an EU intervention to the changes observed' (European Commission, 2017a, p. 57). The incoherence within the Better Regulation Guidelines is clear: causal evaluations are presented as a favoured convention which is difficult to apply to the real world. Since the 1980s and 1990s, evaluation scholarship and practice have increasingly recognised the challenges attached to the use of traditional 'causal evaluation' and have accepted new claims, relating to: 1) the adequacy and credibility of mixed methods

approaches (see e.g. Greene, Caracelli, & Graham, 1989; Mertens & Hesse-Biber, 2013); 2) the need for methodological diversity (European Evaluation Society, 2007); and 3) the invasion of the evaluation literature with the recognition of complexity and how to deal with it (Verweij & Gerrits, 2014).

Rationalist accounts of evaluation can be usefully contrasted with critical approaches, including (but not limited to) those from feminist evaluation scholars. Critical accounts have argued that the strong formalization of evaluation practices enables 'finding use' but impedes 'process use' (the latter defined as 'evaluation use during the process of evaluation; typically use of preliminary results' (Højlund, 2014b, p. 432)); and that the policy-cycle approach results in evaluative '"decision points" every seventh year in the programming phase', preventing a more flexible use of evaluation efforts (Højlund, 2014b, p. 429). Feminist scholars understand evaluation as a tool to advance a particular, substantive agenda; namely gender equality as an agenda of social change. Therefore, social agendas are recognised as complex objects of evaluation (Bustelo, 2003, 2017; Espinosa, 2013; Howard, 2002), notably because social change is slow and requires holistic approaches. As such, it is not realistic to seek direct causal links between single programmes or interventions and observed outcomes.

These critical approaches resonate with complexity theories, which are enjoying a well-deserved revival to theorise the complex systems, contexts and situations with which public policies (and hence evaluation) have to contend (see e.g. Marra, 2015; Walby, 2007). A basic tenet of complexity theory is that systems interact with the environment, thus boundaries (i.e. who and what is considered) matter. Also, the components of systems are interdependent, meaning that interactions are important; there is no-linearity and indeterminacy, resulting in uncertainty; different perspectives shape systems differently through interactions; and processes are outcomes in and of themselves.

Solutions to address the challenges exposed by complexity have been offered by both evaluation scholars and practitioners (see e.g. Espinosa, 2013): mixed method approaches, combining quantitative with qualitative methods to ensure 'a more complex analysis of the social change analysed' (Espinosa, 2013, p. 9); the use of sex-disaggregated data and gender-sensitive analysis throughout the evaluation process (Bustelo, 2017); the use of participatory techniques; a focus on process, i.e. ensuring that prerequisites or impact drivers are in place

(Mergaert & Wuiame, 2013); identifying the difficulties in implementation and addressing these; and, a longitudinal perspective, for quantitative analyses, but also to assess policy learning over time.

Assessing the Commission's rationalist model of evaluation against the requirements of social policy evaluations exposes several weaknesses, three of which have been selected as particularly important. The first weakness is its near exclusive focus on results or outcomes, which excludes attention to the mechanisms in place for addressing social issues (which is necessary for social policy evaluations). The second weakness is the emphasis on quantitative measurement as opposed to a mixed-methods approach. The emphasis on quantitative measurements comes with a failure to recognise the complexity of processes (Caffrey & Munro, 2017) (with this complexity being an inherent part of social policy evaluation). Furthermore, a sharp focus on checking 'what works' has led administrations to 'preoccupations with measurement, traditional worries regarding reliability and validity, and other concerns captured within quantitative methodology' (Shaw, 1999, p. 3). These preoccupations risk detaching the evaluation process from its primary objective. The final weakness is the preference for a single cycle, closed-loop analysis, as opposed to a longitudinal analysis (required to capture more gradual, long-term change). These weaknesses constitute a significant critique of the Commission's dominant evaluation model for assessing gender mainstreaming in EU policies.

2) Evaluating gender mainstreaming in EU research policy in practice

A two-pronged analytical framework is used to explore the EC's approach to evaluation in practice, specifically looking at those evaluations that have taken place across the EU's research framework programmes from FP5 to Horizon 2020.[i] Data was collected from policy texts, including the secondary legislation establishing the Framework Programmes, the evaluation texts themselves (including supporting documentation, e.g. Terms of Reference) and the texts comprising the EC's Better Regulation Agenda. The first part of analysis (section 3.i) considers the individual evaluations, with reference to the evaluation type (i.e. the focus of the evaluation), the actors involved, and the tools used to undertake the evaluations (including data). The second part (section 3.ii) comprises an overarching analysis across the evaluations.

This consists of updating existing longitudinal research that has tracked the consideration of gender equality objectives through evaluation exercises (Mergaert & Minto, 2015), and then interrogating what overarching framework exists for evaluations.

i) Evaluation exercises within EU research policy: types, actors and tools

Our analysis captures 13 different evaluation exercises (text #1 to text #13, from FP5 to Horizon 2020), each of which has resulted in an evaluation text. These exercises were both general (overall evaluations of the framework programme) as well as gender-specific, with eight of the former and five of the latter. The relatively balanced ratio between these two types highlights the resources invested in gender equality and is indicative of the commitments made to gender mainstreaming in EU research policy.

General evaluations serve to verify whether and to what extent the programme delivers its objectives, according to its intervention logic. These evaluations, which are legal requirements,

> 'inform the European Parliament, the Council, Member States, the research community, the public and other stakeholders on: the achievements of the objectives of Horizon 2020 and the continued relevance of all related measures; the programme's efficiency and use of resources, including cross-cutting issues; [and] its EU added-value'. (European Commission, 2017c)

Therefore, they are instruments of accountability, to demonstrate that public money has been well spent. This focus on outcome, relevance and 'value for money' is consistent with the weakness identified in section 1, coming at the expense of more focused attention to process and learning.

Gender-specific evaluation exercises are not legally prescribed. Instead, they are initiated by the unit with responsibility for gender mainstreaming in research policy. Contrary to general evaluations, gender-specific exercises are not primarily preoccupied with measuring progress but are also geared towards learning from what was done. However, while these reports should feed into the general evaluation exercises, several factors (including timing issues) often prevent this (Mergaert & Minto, 2015).

The actors involved and the tools used in the evaluations have varied considerably. Evaluations have been approached in a variety of ways; with some more proceduralised than others. For example, ex-ante evaluations (the IIAs) take place according to a specific set of procedural guidelines (not under investigation here). The approach to ex-post evaluations, however, has evolved over time. Until FP7, interim and ex-post (general) evaluations were performed by panels of selected external 'high level' experts. These panels worked under the supervision of the Commission, drawing principally on Commission-provided sources, such as statistical data and study reports. Notably, the composition of these panels varied between evaluations (without consistent attention to ensuring the panel's gender expertise). The panel would publish an 'external evaluation report', to which the Commission would respond in an accompanying Communication.

The most recent approach for the interim evaluation of Horizon 2020 was markedly different, however. This latest exercise was 'coordinated by the Evaluation Unit of the Commission's Directorate-General for Research & Innovation, with the support of a Working Group and an Inter-Service Group comprising other Commission services' (text #12, page 29). It drew on an extensive data set from a wide variety of sources (underpinned by many different methodologies), including data from statistical analyses, stakeholder consultations, studies by external contractors, and the work of expert groups and panels.

Looking at the actors involved and tools used, gender-specific evaluation-related exercises have been run, and their reports authored, by a range of different actors: external contractors, (external) expert panels, and Commission staff. There is no obvious rationale guiding which approach is adopted. Whereas the work of both external contractors and panels of experts is guided by Terms of Reference, there is greater flexibility in Commission-run exercises. Indeed, texts authored by the Commission have been more restricted in their scope and level of criticism (text #5), and broader in scope and degree of self-criticism (text #9); clearly highlighting the dependence of gender mainstreaming on both the expertise of the Commission officials and the dominant priorities at the time, even within these gender-specific exercises. The availability and use of data are considered further in the section below.

ii) Longitudinal analysis of evaluation practice: tracking objectives and identifying frameworks over time

The second part of the analysis takes a longitudinal view of evaluation in EU research policy. Firstly, research tracks how gender equality objectives have been considered across the evaluations (from FP5 to Horizon 2020), and then it explores the organising framework for guiding evaluations as regards how they promote gender equality.

Tracking gender equality across evaluations

Looking first at the attention paid to gender equality over time, during the period from FP5 to Horizon 2020, there were three gender equality objectives within EU research policy: 1) increasing the number of women in science; 2) increasing the representation of women in decision-making on research activities; and 3) ensuring the needs of both women and men are met through scientific research, i.e. through making research gender-sensitive. A longitudinal analysis from FP5 to Horizon 2020 highlights that these three objectives were not all analysed in every evaluation exercise; with only women in science being more consistently addressed.

Various factors explain this finding. The first (consistent with the findings from section 2) is the relative ease of monitoring the women in science objective using (favoured) quantitative methods. In contrast, more sophisticated (and time-consuming) mixed methods approaches are required to track gender-sensitivity of research. This raises the issue of data more broadly, which continues to hinder effective evaluation. The Interim Evaluation of Horizon 2020 (text #12) noted that there are problems with the quality and comparability of quantitative data, which complicates analyses and evaluation work. This issue was highlighted in several gender-specific and general evaluations, although there has been no remediation to date. For example, the criticism features in text #7, as well as the ex post evaluation report of FP7 (text #11), the latter of which states: 'it is impossible to analyse gender issues to a satisfying degree. While detailed information about female participation and gender issues was collected systematically throughout FP6, such information was no longer collected for FP7.' (p.38). This emphasises the lack of any consistent approach to the collection of robust and comparative data over time. Despite the lack of relevant data, this latter report (text #11) addressed gender equality in FP7 to some extent. It considered process-related aspects (e.g. noting that the counting of projects integrating gender issues in the content of their work was based on simple self-reporting by

project holders, in a 'tick-the-box' form) and result-related aspects. The availability of gender expertise on the panel (by way of a gender and evaluation expert) would have supported this more comprehensive approach to considering gender equality, notwithstanding the lack of data.

The second and related explanatory factor is the failure to integrate all evaluation exercises. Over this period of analysis, gender-specific evaluations were not fully reflected within 'mainstream' (or general) evaluations, with political and institutional factors working against the inclusion of gender-specific evaluations (Mergaert & Minto, 2015). The timing of evaluations was seen to work against effective integration, as was the case with the latest (general) interim evaluation of Horizon 2020 (text #12). The list of 'input studies' for this evaluation consists of 38 different items; however, the report of the dedicated evaluation of 'Gender equality as a crosscutting issue in Horizon 2020' (text #13) was published too late to serve as a source, undermining effective gender mainstreaming. Indeed, the report offered a much more fine-grained gender-sensitive analysis, and a series of recommendations, organised per phase of the Horizon 2020 process and for each of the three gender equality objectives. In a bid to secure some gender-specific input into the interim evaluation, the Expert Group for the gender-specific evaluation was tasked to deliver, only ten weeks after its launch (including the Christmas break), a five-page report for the High-Level Expert Group undertaking the general interim evaluation. At such a nascent stage in the process, the five-page report had a limited focus on some aspects of the Horizon 2020 gender equality strategy and intervention logic. When the gender-specific evaluation report was complete four months later (text #13), the High-Level Expert Group for the general evaluation had already finished its report (text #12). The report from the High Level Expert Group boasts that the programme 'leads by example in gender' (text #12, page 126) and that 'progress is made with respect to promoting gender equality under Horizon 2020' (page 195). These claims are poorly founded.

Frameworks for evaluation

Regarding the overarching structure within which evaluations take place, the general evaluation exercises are mandated within the directives that establish the framework programmes and/or within the Commission's established governance practice. However, this proceduralisation is only seen at the level of the framework programmes. There is no clear, overarching framework for evaluations in EU research policy that takes a longitudinal perspective across framework programmes. Therefore, aside from the five-year assessments

(which were discontinued in 2003), evaluations are organised around a single policy cycle (which is consistent with the findings from section 2). This results in ex-ante, interim and ex-post evaluations of each framework programme that are not embedded within a longer-term or cross-framework analysis.

In the framework of its Better Regulation Agenda, the EC developed specific Guidelines (European Commission, 2017a) and a Toolbox (European Commission, 2017b) that span the policy cycle and thus integrate impact assessment, monitoring and evaluation (including fitness checks). As these apply to all evaluation exercises, this is the closest likeness that the Commission has to an overarching framework for evaluation. However, as a framework to support the evaluation of gender mainstreaming, it is lacking. The principal flaw of the Guidelines and Toolbox is that they concern the evaluation of interventions themselves, without addressing cross-cutting issues within these interventions. Therefore, they have no direct bearing on the (effective) evaluation of gender mainstreaming in EU policy-making. At best, they could set a framework for evaluating the gendered-impacts of a particular intervention. However, even on this count, the Better Regulation Toolbox is weak, with gender included as *part* of one of six non-mandatory criteria to be addressed within an evaluation, specifically "Equity: how fairly are the different effects distributed across the different stakeholders / regions? / genders? / Social groups?" (p. 355).

Neither do the Better Regulation Guidelines refer to gender equality explicitly. Rather, gender equality appears subsumed under the broader umbrella of social concerns. The Guidelines state that among the key requirements for evaluations (including fitness checks): "All evaluations must assess all significant economic, social and environmental impacts of EU interventions (with particular emphasis on those identified in a previous [IIA]) or explain why an exception has been made" (p.50). However, when referring to the Better Regulation Toolbox (Chapter 3), we see that gender equality is not listed under social impacts. Instead it is included as a Fundamental Rights consideration, to be assessed across the economic, social and environmental impacts (p. 125), with a list of associated questions regarding the impact (direct and indirect) on women, men and (the promotion of) gender equality (p.133). Whilst not insignificant, the corresponding Tool #28 on Fundamental Rights and Human Rights (pp.209-212) makes no explicit reference to either gender equality or even to (the narrower conceptualisation of) women's rights. Therefore, there is an incoherence between the treatment

of gender equality in the Better Regulation Guidelines on the one hand (where it is addressed as a social issue) and the Toolbox on the other (where it is addressed as a cross-cutting fundamental rights issue); and there is also a gap within the Toolbox itself, with gender missing at the level of the tools. The incoherence and absence with respect to gender within these key governance texts does little to ensure that gender mainstreaming is addressed consistently in all evaluations.

The lack of an overarching evaluation strategy with respect to gender equality partly explains why there is little continuity in policy objectives between evaluations and such an eclectic range of tools and data employed. Without a longitudinal framework, there is no consistent approach to evaluating gender mainstreaming laid down in the Terms of Reference for the evaluation exercises and neither is gender equality protected within the evaluation process. In the context of larger and increasingly complex framework programmes, which necessitate hugely complex evaluation processes, the three gender equality objectives are rendered acutely vulnerable to exclusion if there is no requirement for their inclusion. In the most recent general evaluation (Interim Evaluation of Horizon 2020), as noted in the report itself, the list of cross-cutting issues for consideration includes, "sustainable development, climate and biodiversity action, more responsible R&I, gender equality, integration of social sciences and humanities (SSH) in R&I projects, and generating outputs for policy" (text #12, p.159). It is notable that gender equality features; however, there is a risk that it is side-lined or 'evaporates' (European Commission, 2016, p. 2) in such a complex process that places no obligation upon the evaluators. Indeed, looking at the Terms of Reference, gender is mentioned once: 'The interim evaluation shall also take into consideration the scope for further simplification and aspects relating to access to funding opportunities for participants in all regions and for the private sector, notably SMEs, as well as the scope for promoting gender balance.' (European Commission, 2016, p. 2) . The focus on 'gender balance' reduces the threefold gender equality objective in EU research policy to its quantitative dimension, overlooking the objective to ensure gender-sensitive research.

iii) Conceptual and procedural weaknesses in the Commission's approach

Our analyses identify the same three weaknesses in the Commission's approach in theory (section 2) and in practice (sections 3.i and 3.ii): results over process; quantitative over mixed

methods; and single cycle over longitudinal analysis. However, the empirical analysis provides a more nuanced understanding of how the weaknesses, as they appear from the theory, play out in practice. Looking at each of the weaknesses in turn, the first can be related to the purpose of evaluation as, used as an accountability device (without emphasis upon policy learning potential), evaluation exercises will prioritise measurement of results over an assessment of the process. The second weakness is the preference for quantitative over mixed methods analysis. This is shaped by practical (as well as more principled) factors, which require the condensing, synthesis and analysis of this complex data as part of the evaluation process. This may lead to a preference for quantitative data, given the relative ease of data collection, manipulation and communication in comparison to a mixed-methods approach. Indeed, even in cases where an evaluation exercise does have access to a range of varied data sources, ultimately these must be consolidated into a discrete set of conclusions and recommendations. In addition to the type of data preferred is the matter of the availability of data and its quality. The unavailability of quality data was specifically highlighted as a barrier to the evaluation of gender mainstreaming and certainly precludes the exploration of change (or 'added value') over the long-term. These points also resonate with the final weakness identified in theory, which is the preference for a closed-loop (as opposed to longitudinal) approach. Without such an approach, the integration of gender-specific exercises within general evaluations is undermined and does not support a coherent, longitudinal approach where there is consistent attention to gender equality.

Beyond these three weaknesses in the EC's approach to evaluation, the investigation of evaluation in practice has highlighted other underlying factors that impact on the evaluation of gender mainstreaming in EU research policy; namely, expertise, agency, timing and data. Taking each in turn, given the complexity of evaluating programmes, when constructing an evaluation panel (either internal or external) expertise in evaluation itself is essential. Furthermore, gender expertise is a prerequisite for the evaluation of gender mainstreaming, to enable a full consideration of the multiple and complex dimensions of gender equality. There is little evidence of consistency across the evaluation exercises with respect to the representation of gender expertise and how that expertise is used within the evaluation process. The issue of expertise is linked closely to the second factor, which is agency. Without a requirement to address all gender equality objectives, it is left to the discretion of the actors involved as to what will be included within and excluded from the evaluation. Of course, this can work both for and against attention to gender equality. Gender experts and femocrats may

be able to exploit opportunities to successfully push for greater attention to gender; however, there is a risk that gender concerns fall off the agenda when actors do not have either gender expertise or feminist sympathies.

Beyond the actors involved in the evaluation process, as a third point, the timing of the evaluations impacts on the effective evaluation of gender mainstreaming (see also Mergaert & Minto, 2015), which precludes the inclusion of results from gender-specific evaluations within mainstream exercises. Finally, various issues relating to data are a concern for the effective evaluation of gender mainstreaming. Beyond the availability and use of high-quality data from mixed methods (discussed above) is the matter of the sheer volume and complexity of the data sources. For example, there were 38 input studies provided for the Interim Evaluation of Horizon 2020 (text #13), excluding the gender-specific evaluation text.

Conclusions and way forward

Situated within the expanding field of post-adoption feminist scholarship, through this research we have sought to explore the suitability of the Commission's dominant approach to evaluation to adequately support the promotion of gender mainstreaming. In so doing, we make an original contribution to a particularly under-explored element of this post-adoption feminist research, namely policy evaluation. As with other social objectives, the promotion of gender equality poses particular challenges for evaluation practice and, acknowledging this, corresponding evaluation approaches have been advanced by scholars and practitioners. We have investigated the extent to which the requirements of these agendas are reflected in the Commission's approach to evaluation, in both theory and practice, taking gender mainstreaming in EU research policy as a hypothesis-generating case study (Lijphart, 1971).

Analysis of evaluation in both theory (section 2) and practice (section 3) exposed three particular weaknesses in the Commission's approach, specifically: 1) the rationalist focus on results or outcome, to the exclusion or marginalisation of process (the latter being an essential part of social policy evaluation); 2) the emphasis on quantitative measurement as opposed to a mixed methods approach (a necessary element of social policy evaluation); and 3) the preference for a single cycle, closed-loop analysis, as opposed to a longitudinal analysis

(required to capture more gradual, long-term change). Our empirical analysis of the evaluation exercises from FP5 to Horizon 2020 highlighted additional factors for consideration; namely expertise, agency, timing and data. Taken together, these findings provide a rich account of where the Commission's approach falls short in the promotion of gender equality (and other complex social agendas) through the evaluation process.

Exploring each weakness in turn, the first relates to the prioritisation of evaluation's accountability role, despite the well-established agreement amongst scholars and practitioners that the learning role must be emphasised if evaluation policies are to contribute to policy advancement. By the same token, the second weakness relates to another widely supported facet of evaluation practice, that of mixed methods. Unfortunately, mixed methods continue to suffer from strong epistemological resistance, as the qualitative element is often considered as a poorer but necessary replacement when the "real" (gold standard) evidence is not available. The call for condensed and supposedly easy-to-grasp data supports the development of simplistic evaluative solutions to complex realities which, ultimately, have little value (if any at all). The third weakness, regarding the lack of a longitudinal perspective, undermines the enlightenment purpose of evaluation; that is, its capacity for policy learning and policy improvement which are the ultimate goals for a useful evaluation. Finally, relating to the additional identified factors for consideration, whilst expertise and agency have a specific importance for gender-responsive and feminist approaches to evaluation, the importance of appropriate timing and the availability of data are more generally applicable as basic tenants of any robust evaluation policy.

The particular contribution of this paper lies in its in-depth illustration of the Commission's approach to evaluation, in theory and in practice. This is an essential first step for the development of more robust hypotheses for feminist and gender-responsive analyses of policy evaluation. The research agenda in this area must acknowledge that the Commission is not a uniform institution. Instead, there are different institutional and evaluation sub-cultures within the different DGs; cultures which will be more or less accommodating of gender mainstreaming and other cross-cutting social agendas. Therefore, analyses across policy areas are essential. Furthermore, whilst we argue that the findings from this research reflect a broader social policy agenda (beyond gender equality), more focused analyses into different social agendas would lend further nuance to our findings. As a third thread of a research agenda, the concept of

resistance merits additional research. Notwithstanding differentiation across the Commission, its dominant approach to evaluation has not mainstreamed gender or other cross-cutting agendas (despite their constitutionalised status and their inclusion at a policy and programme level). Understanding the dynamics of the institutional and individual resistances to gender mainstreaming in evaluation would serve to enhance our responses to correct and reform these processes.

As prescribed in Article 8 TFEU, gender equality should be mainstreamed in all stages of the policy cycle, including policy evaluation. The following recommendations strengthen current evaluation practice to better contribute to the promotion of the EU's cross-cutting social agendas. Adopting a longitudinal approach to the evaluation of gender mainstreaming in EU research policy is a necessary step; fixing gender equality objectives across framework programmes. As part of this, provision for reliable and comparable mixed methods data to monitor these objectives is essential. Finally, the inclusion of gender expertise must be a fundamental requirement for all evaluations, including general evaluations undertaken external to the Commission.

Bibliography

Abels, G. (2012). Research by, for and about Women: Gendering Science and Research Policy. In G. Abels & J. M. Mushaben (Eds.), *Gendering the European Union: New Approaches to Old Democratic Deficits* (pp. 187-207). Basingstoke: Palgrave Macmillan.

Allwood, G. (2013). *Gender mainstreaming and policy coherence for development: Unintended gender consequences and EU policy.* Paper presented at the Women's Studies International Forum.

Braithwaite, M. (2000). *Mainstreaming Gender in the European Structural Funds*. Paper presented at the Mainstreaming Gender in European Public Policy Workshop, University of Winconsin-Madison. http://eucenter.wisc.edu/Conferences/Gender/braith.htm

Bustelo, M. (2003). Evaluation of gender mainstreaming: Ideas from a meta-evaluation study. *Evaluation, 9*(4), 383-403.

Bustelo, M. (2017). Evaluation from a Gender+ Perspective as a Key Element for (Re)gendering the Policymaking Process. *Journal of Women, Politics & Policy, 38*(1), 84-101. doi:10.1080/1554477X.2016.1198211

Bustelo, M., & Verloo, M. (2009). Grounding policy evaluation in a discursive understanding of politics. In.

Caffrey, L., & Munro, E. (2017). A systems approach to policy evaluation. *Evaluation, 23*(4), 463-478. doi:10.1177/1356389017730727

Callerstig, A.-C. (2014). Can public procurement be an instrument for policy learning in gender mainstreaming? *Offentlig förvaltning. Scandinavian Journal of Public Administration, 18*(4), 51-71.

Cavaghan, R. (2013). Gender mainstreaming in the DGR as a knowledge process: epistemic barriers to eradicating gender bias. *Critical Policy Studies, 7*(4), 407-421.
Cavaghan, R. (2017). *Making gender equality happen: Knowledge, change and resistance in EU gender mainstreaming*: Routledge.
Cavaghan, R. M. (2012). *Gender Mainstreaming as a Knowledge Process: towards an understanding of perpetuation and change in gender blindness and gender bias.* Edinburgh:[Sn],
de Gregorio Hurtado, S. (2017). A critical approach to EU urban policy from the viewpoint of gender. *Journal of Research in Gender Studies, 7*(1).
Debusscher, P. (2011). *Mainstreaming gender in European commission development policy: conservative Europeanness?* Paper presented at the Women's Studies International Forum.
Debusscher, P., & Van der Vleuten, A. (2012). Mainstreaming gender in European Union development cooperation with sub-Saharan Africa: Promising numbers, narrow contents, telling silences. *International Development Planning Review, 34*(3), 319-338.
Engeli, I., & Mazur, A. (2018). Taking implementation seriously in assessing success: the politics of gender equality policy. *European Journal of Politics and Gender, 1*(1-2), 111-129.
Espinosa, J. (2013). *Promoting human rights and gender sensitive evaluations: key ideas for evaluating gender equality results*.
European Commission. (1996). *Communication on evaluation, concrete steps towards best practice across the commission (Document No. SEC 96/659)*. Retrieved from Brussels:
European Commission. (2010). *Smart Regulation in the European Union, Communication, COM(2010) 543 final, 8 October 2010*.
European Commission. (2012). *A Reinforced European Research Area Partnership for Excellence and Growth, Communication, COM(2012) 392 final, 17 July 2012* Brussels
European Commission. (2013). *Strengthening the foundations of Smart Regulation – improving evaluation*. Brussels, 2 October 2013
European Commission. (2016). *Terms of Reference for the Interim Evaluation of Horizon 2020, approved 2 September 2016*. Brussels
European Commission. (2017a). *Better Regulation Guidelines, Staff Working Document, SWD (2017) 350, 7 July 2017*. Retrieved from Brussels:
European Commission. (2017b). *Better Regulation Toolbox (complementing the Better Regulation Guidelines, SWD(2017) 350))*. Retrieved from https://ec.europa.eu/info/sites/info/files/better-regulation-toolbox_1.pdf
European Commission. (2017c). 'Research and Innovation - Evaluations' Webpage, last accessed 5 January 2018. Retrieved from https://ec.europa.eu/research/evaluations/index_en.cfm
European Evaluation Society. (2007). *EES Statement: The importance of a methodologically diverse approach to impact evaluation - specifically with respect to development aid and development interventions, December 2007*. Retrieved from https://www.europeanevaluation.org/sites/default/files/EES%20Statement_0.pdf
Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis, 11*(3), 255-274.
Højlund, S. (2014a). *Evaluation in the European Commission - For accountability or learning?* Paper presented at the Jean Monnet Workshop on Policy Evaluation in the European Union, Cardiff Jean Monnet Centre.
Højlund, S. (2014b). Evaluation use in evaluation systems – the case of the European Commission. *Evaluation, 20*(4), 428–446.
Howard, P. L. (2002). Beyond the 'grim resisters': Towards more effective gender mainstreaming through stakeholder participation. *Development in Practice, 12*(2), 164-176. doi:10.1080/09614520220127685
Hubert, A. (2012). Gendering employment policy: from equal pay to work-life balance. In *Gendering the European Union* (pp. 146-168): Springer.

Krizsan, A., & Roggeband, C. (2018). Towards a conceptual framework for struggles over democracy in backsliding states: Gender equality policy in Central Eastern Europe. *Politics and Governance, 6*(3), 90-100.
Kronsell, A. (2012). Gendering theories of European integration. In *Gendering the European Union* (pp. 23-40): Springer.
Kronsell, A. (2015). The Power of EU Masculinities: A Feminist Contribution to European Integration Theory. *JCMS: Journal of Common Market Studies*.
Lijphart, A. (1971). Comparative Politics and the Comparative Method. *The American Political Science Review, 65*(3), 682-693.
Linková, M., & Červinková, A. (2011). What matters to women in science? Gender, power and bureaucracy. *European Journal of Women's Studies, 18*(3), 215-230.
Marra, M. (2015). Cooperating for a more egalitarian society: Complexity theory to evaluate gender equity. *Evaluation, 21*(1), 32-46.
Mazur, A. G. (2017). Toward the systematic study of feminist policy in practice: An essential first step. *Journal of Women, Politics & Policy, 38*(1), 64-83.
Mergaert, L. (2012). *The Reality of Gender Mainstreaming Implementation. The Case of the EU Research Policy.* (Doctoral dissertation), Radboud Universiteit Nijmegen, Nijmegen.
Mergaert, L., & Demuynck, K. (2011). *The ups and downs of gender mainstreaming in the EU research policy - the gender toolkit and training activities in FP7*. Antwerp, Belgium: Policy Research Centre on Equal Opportunities.
Mergaert, L., & Minto, R. (2015). Ex *ante a*nd ex *post e*valuations: two sides of the same coin? The case of gender mainstreaming in EU Research Policy. *European Journal of Risk Regulation, 1*, 47-56.
Mergaert, L., & Wuiame, N. (2013). *Report on Institutional Capacity for Gender Mainstreaming in the European Commission*. Retrieved from Report from a study for the European Institute for Gender Equality (unpublished work):
Mertens, D. M., & Hesse-Biber, S. (2013). Mixed Methods and Credibility of Evidence in Evaluation. *New Directions for Evaluation, 2013*(138), 5-13. doi:10.1002/ev.20053
Minto, R., & Mergaert, L. (2018). Gender mainstreaming and evaluation in the EU: comparative perspectives from feminist institutionalism. *International Feminist Journal of Politics, 20*(2), 204-220. doi:10.1080/14616742.2018.1440181
Moser, C. (2005). Has gender mainstreaming failed? A comment on international development agency experiences in the South. *International Feminist Journal of Politics, 7*(4), 576-590.
Pollack, M., & Hafner-Burton, E. (2000). Mainstreaming Gender in the European Union. *Journal of European Public Policy, 7*(3), 114-138.
Prügl, E. (2012). The common agricultural policy and gender equality. In *Gendering the European Union* (pp. 127-145): Springer.
Shaw, I. (1999). *Qualitative Evaluation*. London: Sage.
Stern, E. (2009). Evaluation policy in the European Union and its institutions. *Evaluation policy and evaluation practice. New Directions for Evaluation, W.M.K. Trochim, M. M. Mark & L. J. Cooksy (eds), 123*, 67–85.
Summa, H., & Toulemonde, J. (2002). Evaluation in the European Union: Addressing Complexity and Ambiguity. In J.-E. Furubo, R. C. Rist, & R. Sandahl (Eds.), *International Atlas of Evaluation* (pp. 407-425). New Brunswick and London: Transaction Publishers.
True, J. (2009). Trading-in gender equality: Gendered meanings in EU trade policy. In *The discursive politics of gender equality* (pp. 141-157): Routledge.
Van der Vleuten, A. (2016). *The price of gender equality: Member states and governance in the European Union*: Routledge.
Verloo, M. (2018). *Varieties of opposition to gender equality in Europe*: Routledge.
Versluis, E., van Keulen, M., & Stephenson, P. (2011). *Analyzing the European Union Policy Process* Basingstoke, New York: Palgrave Macmillan.

Verweij, S., & Gerrits, L. (2014). Managing unplanned events in large infrastructure projects: Results from an in-depth comparative case evaluation. *Compact II: Administrative strategies for complex governance systems*, 81-108.
Walby, S. (2007). Complexity theory, systems theory, and multiple intersecting social inequalities. *Philosophy of the social sciences, 37*(4), 449-470.

Funding details

The research of María Bustelo has been funded by the Spanish Ministry of Science, Innovation and Universities UNIGUAL Project Gender Equality Policies in Spanish University (Ref: Fem2017-86004-R).

Conflict of interest statement

The Authors declare that there is no conflict of interest

Acknowledgements

The authors would like to thank the editor, Prof. Isabelle Engeli, and the anonymous reviewers for their valuable contributions to development of this article.

Author biography

Rachel Minto is a Research Fellow at Cardiff University's Wales Governance Centre.

Lut Mergaert is Director and in charge of policy design at Yellow Window, a Belgium-based consultancy.

María Bustelo is Associate Professor of Political Science and Public Administration at Complutense University of Madrid

[i] **Text 1 (2000):** Five-Year Assessment of the European Union Research and Technological Development Programmes, 1995-1999, Report of the Independent Expert Panel chaired by Joan Majó

Text 2 (2001): Commission of the European Communities (2001) "Gender in Research. Gender Impact Assessment of the specific programmes of the Fifth Framework Programme. An overview." Undertaken by external contractors

Text 3 (2004): Five-Year Assessment of the European Union Research and Technological Development Programmes, 1999-2003, Full Report, prepared by an external expert group

Text 4 (2005): Commission of the European Communities (2005) "Annex to the Proposal for the Council and European Parliament decisions on the 7th Framework Programme (EC and Euratom). Main Report: Overall summary Impact assessment and ex ante evaluation", 6 April 2005, SEC(2005) 430

Text 5 (2008): Commission of the European Communities (2008) "Gender equality report. Sixth Framework Programme." October 2008, undertaken by the Commission

Text 6 (2009): Expert Group on the *ex post* evaluation of the sixth Framework Programmes (2002-2006) (2009) "Evaluation of the Sixth Framework Programmes for Research and Technological Development 2002-2006"

Text 7 (2009): European Commission (2009) "Monitoring progress towards Gender Equality in the Sixth Framework Programme, Synthesis Report." May 2009, compiled by the Centre for Strategies and Evaluation Services (CSES) on the basis of contributions by the Gender Monitoring Studies Contractors and the European Commission

Text 8 (2010): Interim Evaluation of the Seventh Framework Programme (2007-13), Report of the Expert Group, 12 November 2010, Chair: Rolf Annerberg

Text 9 (2010): European Commission (2010) "Stocktaking 10 years of "Women in Science" policy by the European Commission 1999-2009"

Text 10 (2011): European Commission (2011) "Horizon 2020 - The Framework Programme for Research and Innovation - Impact Assessment Report", 30 November 2011, SEC(2011) 1427 final

Text 11 (2015): Commitment and Coherence - Ex-Post-Evaluation of the 7th EU Framework Programme (2007-2013), Report from the High Level Expert Group, Chair: Louise O. Fresco

Text 12 (2017): European Commission, Interim evaluation of Horizon 2020, Commission Staff Working Document

Text 13 (2017): Interim Evaluation: Gender equality as a crosscutting issue in Horizon 2020, September 2017, Report of the Expert Group, Chair: Suzanne de Cheveigné